A PROV standard-based data source agnostic provenance engine for Big Data analytics

CASE WESTERN RESERVE UNIVERSITY

PI: SAHOO, SATYA SANKET

Grant Number: 1 U01 EB020955-01

Data provenance is key to ensuring data quality, scientific reproducibility, and tracing the lineage of data as it undergoes transformation for use n the ""data-driven"" research paradigm. The emerging ""Big Data"" resources in biomedical research and clinical care domains have highlighted multiple computational challenges to develop a scalable and high performance provenance analysis engine. These computational challenges include semantic heterogeneity across provenance information generated from disparate sources (variety), lack of scalable provenance analytical algorithms that can keep pace with large volume of data generated at a rapid velocity. Using the new PROV representation standard recommended the W3C, which is the standard body for Web technologies, together with distributed cloud computing technologies we propose to develop a highly scalable data source agnostic provenance engine. To address the lack of appropriate provenance analytical operations required to develop this provenance engine over the PROV representation model, we will follow a three-phase approach: (1) we will first develop a new algebraic graph framework for analyzing provenance graphs conforming to the PROV standard, (2) in the second phase we will use the insights from the systematic characterization of provenance analysis operations to define distributed algorithms for implementation over cloud computing technologies, and (3) in the final step, we will implement the provenance engine that will support three fundamental provenance functions of (a) scientific reproducibility, (b) data quality assurance, and (c) trust computation. The resulting provenance engine will potentially transform the use of provenance in biomedical ""Big Data"" exploration and analysis techniques in the increasing number of data repositories such as the National Sleep Research Resource for accelerating data-driven research in disease mechanisms. PUBLIC HEALTH RELEVANCE PUBLIC HEALTH RELEVANCE: Data provenance is key to ensuring data quality, scientific reproducibility, and tracing the lineage of data for use in the ""data-driven"" research paradigm. The goal of this project is to use the PROV provenance model together with distributed cloud computing technologies to develop a data source agnostic provenance engine. Results from this project enable acceleration of research in disease mechanisms through biomedical ""Big Data"" analytics.